长期天气预报线性自回归模式的拟合

么 枕 生

(南京大学气象系)

编者按：本文总结了平稳时间序列线性自回归模式拟合的各种方法。线性自回归模式的参数是由Durbin的逐步递推过程进行估计的，该模式的阶数通过t检验或F检验予以选定。文中还给出用于月降水量预报的说明性实例。

1. 引 言

观测的时间序列可作为是随机过程的一种特殊实现。随机模式是随机过程的公式化表达。随机模式，与自回归模式和马尔柯夫链模式，对于气象学是十分有用的。

状态空间为离散的马尔柯夫过程，称为马尔柯夫链。Gabriel及Neumann (1962) 首先提出将一阶马尔柯夫链用于研究以色列的Tel Aviv城的日降水量问题。用以拟合Tel Aviv日降水量资料的马尔柯夫链的阶数，是有争议的(Gates及Tong，1976;Katz,1979)。为了检验马尔柯夫链模式拟合于日降水量记录的优度，Lowry及Guthrie (1968) 根据Billingsley的理论(1961)及Hoel的工作(1954)提出一种估计马尔柯夫链模式阶数的χ^2检验。根据Akaike (赤池) 信息判据(Akaike，1974)，Gates及Tong(1976)接着Tong(1975)的工作而提出估计马尔柯夫链模式阶数的一种模式建立程序。利用了几种程序，Katz (1979) 将Akaike信息判据及Bayesian信息判据(Schwarz,1978)用于马尔柯夫链模式阶数的最佳选择，并且对这两种程序进行了比较。然而，为了改进马尔柯夫最小反馈特性在研究湿日和干日概率的变化，作者(1966)曾将天气的随机变化作为试验来处理，并将任何过去试验的影响都纳入现在状态的分析之中。这种处理方法使得在不考虑如何确定马尔柯夫链模式阶数问题的情况下，有可能给出几步转换的正规马尔柯夫链。

本文将讨论有关线性自回归模式拟合于观测序列的问题。p阶的自回归模式简写为AR(p)模(Box及Jenkins,1970)。Yule(1927)发展了经典的周期性概念，认为AR过程可用来表示由一系列随机外部冲击所干扰形成的简谐振荡。这种过程的实现之所以称之为随机的，就是因为未来值是由过去值部分地确定出的。

在将线性AR模式拟合于平稳时间序列的各种方法中，Yule-Walker方程组是大家都知道的。此外，这些有关的参数常常用Durbin(1960)提出的逐步递推过程予以估计。

Box和Jenkines(1970)提出一种供估计这些参数用的系统性方法。他们首先找出得自于Yule-Walker方程组的初始估计量，然后用迭代程序对这些初始量进行调整，以期获得AR模式的最小二乘方估计量。

Mann 及 Wald(1943)曾证明 AR 模式中最小二乘方估计量的抽样特性，如同多元正态情况下最小二乘方回归估计量一样，是渐近的。根据这些有用的定理，我们可以将 AR 模式中的估值问题作为普通的回归问题来处理。于是，Chatfield(1980)指出，高阶 AR 模式可采用相似于包括在拟合普通回归模式中的最小二乘方的办法予以拟合。Yule-Walker 方程组同样可以由最小二乘方方法得出。

Jones(1964)曾采用一种不同的自回归拟合法，他利用得自于周期图的谱来估计相关参数。此周期图为自协方差函数的有限傅里叶变换式。

用于确定线性 AR 模式的检验是多种多样的。Quenouille(1947)曾用偏自相关系数提供了一种机巧的拟合检验（见 Kendall 及 Stuart，1976 b，pp. 502—505）。

按照 Mann-Wald 定理，线性多元回归中的 F 检验可用于选择 AR 模式的阶数(Yevjevich，1972)。Zurndorfer 及 Glahn(1977)在前人研究的基础上进一步考虑了各预报因子间的相关，完成了一种检验回归模式显著性的 F 统计量的 Monte Carlo 方法。这种 Monte Carlo F 值可看成是确定线性 AR 模式的合适阶数(即没有过拟合)的一种新方法。

另一种方法是计算 AR 模式中拟合剩余项时的剩余平方和，当拟合的改进很小时就有可能找出模式的合适阶数(Kendall 及 Stuart，1976；Chatfield，1980)。Yevjevich(1972)断然认为，白噪化方法似乎是确定线性 AR 模式阶数的最有吸引力的方法，如果随机误差序列与由 Yevjevich 意义上的纯随机序列(白噪音序列)没有显著差别，我们虽然可以假定，在此模式[见下面的(1)式]中随机误差序列是一种白噪音序列，但由于它是在物理上不可能实现的现象，所以这样的假定有时却并不合适。正如由 Chatfield(1980)指出的那样，此种结论为建立下面要讨论的模式 III 及 IV 奠定了进一步的基础，尽管这些模式的随机误差序列较接近于近似的白噪音，但是事实上它们却并不合适。

Akaike(1969)曾建议，线性 AR 模式的阶数常可用大家都知道的 Akaike 最终预报误差(FPE)予以选定。另一种鉴定该模式阶数的客观处理方法，是设法减小由 Akaike(1974)提出的 Akaike 信息判据(AIC)的量或所谓的 Bayesian 信息判据(BIC)的量。

Shibata(1980)提出一种选择 AR 模式阶数的渐近有效选择法，该文中将 AIC、BIC 等六种选择阶数的方法进行了比较。

近年来，Carr (1980)导出新判据 L_1 及 L_2 以确定线性多元回归及线性 AR 模式中所包括的预报因子的数目。

这些方法对或者是用于经验性近似关系或者是用于定量判据以选择模式的合适阶数方面，在它们企图在所建立的模式中减小误差及减少项数要求上大都相类似。看来，这些常用的选择某个模式阶数的方法的缺点是，通常都缺乏关于参数统计显著性的说明，并且需要较为费时的计算。

为了拟合观测的时间序列，Box 及 Jenkins (1970) 认为偏自相关函数作为选择 AR 过程（模式)的一种度量标准。按照 Box 及 Jenkins 的意见，如是自相关函数，当 $k \leqslant p$ 时不为零，而当 $k > p$ 时为零，那么此 AR 应该是 p 阶的。Chatfield (1980) 为检验偏自相关给出 95%的 $\pm 2/\sqrt{n}$ 的置信区间。可是，Chatfield 所用的 $\mathrm{var} r_k \approx 1/n$ 只是一种近似结果；这是假设除了后延零的自相关外，所有的总体自相关系数均为零。

本文以 Box 及 Jenkins(1970) 的研究成果为基础，认为线性 AR 模式的阶数能以预选定的显著性水平用 t 检验或 F 检验可恰当地予以选出，这种计算是直截了当的。但应牢记，t 检验和 F 检验只是对正态独立随机变量才能成立。

本文中所用的雨季月降雨量是正态分布的(Brooks 及 Carruthers，1953；Yao，1958，1963)。根据中心极限定理，t 检验或 F 检验所要求的正态条件，对于具有同样分布的独立随机变量的大样本，是得到满足的。进一步说，雨量记录的分布曲线是随观测时期的长度而变化的；例如，小时降雨量有很强的偏倚分布，可是年降雨量却只稍有偏倚(准正态的)。一般说来，雨量的分布随着观测时期的增长而越接近于正态分布(Brooks 及 Carruthers，1953)。去掉偏倚的一种标准的统计方法是采用某种变换（诸如对数、平方根或立方根)将有偏分布的原始观测值变换为新的、具有正态分布的观测值(Katz 及 Skaggs，1981)。就此而论，除了要求变量的独立性外，为 t 检验及 F 检验所要求的正态条件也能由变换技术满足之。

2．线性自回归模式的建立

设 y_t 为平稳随机过程对平均值的离差，则表示平稳时间序列(随机过程)的线性 AR 模式如下：

$$y_t=\alpha_1 y_{t-1}+\alpha_2 y_{t-2}+\cdots+\alpha_p y_{t-p}+\varepsilon_t \quad (1)$$

模式(1)为现在值 y_t 对该过程的过去值 $y_{t-1},y_{t-2},\cdots,y_{t-p}$ 的回归。$\{\varepsilon_t\}$为正态随机变量序列，其中每一个值都具有平均数零和方差 σ_ε^2(白噪音)。这个具有参数 $\alpha_1,\alpha_2\cdots,\alpha_p$ 的模式称为线性AR(p)模式。

首先，通过求解下列方程组

$$\left.\begin{aligned}\rho_1&=\alpha_1+\alpha_2\rho_1+\cdots+\alpha_p\rho_{p-1}\\ \rho_2&=\alpha_1\rho_1+\alpha_2+\cdots+\alpha_p\rho_{p-2}\\ &\cdots\\ \rho_p&=\alpha_1\rho_{p-1}+\alpha_2\rho_{p-2}+\cdots+\alpha_p\end{aligned}\right\} \quad (2)$$

可估计这些线性自回归参数。这是 Yule-Walker 方程组，这组线性方程具有自回归参数 $\alpha_1,\alpha_2,\cdots,\alpha_p$，这些参数都通过自相关系数 $\rho_1,\rho_2\cdots,\rho_p$ 而彼此联系着。

当 Durbin 的逐步递推程序 (Durbin, 1960) 被用作第二种估计线性 AR 模式参数的方法时，递推公式如下：

$$\left.\begin{aligned}\alpha_k^{(k)}&=\frac{\rho_k-(\alpha_1^{(k-1)}\rho_{k-1}+\alpha_2^{(k-1)}\rho_{k-2}+\cdots+\alpha_{k-1}^{(k-1)}\rho_1)}{1-(\alpha_1^{(k-1)}\rho_1+\alpha_2^{(k-1)}\rho_2+\cdots+\alpha_{k-1}^{(k-1)}\rho_{k-1})}\\ \alpha_j^{(k)}&=\alpha_j^{(k-1)}-\alpha_k^{(k)}\alpha_{k-j}^{(k-1)}\\ &j=1,2,\cdots,k-1;k=1,2,\cdots p\end{aligned}\right\} \quad (3)$$

式中上角标(k)表示此模式的阶数，下角标 k 则表示第 k 个参数。于是，$\alpha_k^{(k)}$为 k 阶线性 AR 模式的最后参数。

根据 Mann-Wald 定理，线性 AR 模式的参数可通过偏自相关来估计，所得结果与得自(2)式或(3)式的相同。现在，作为第三种估计线性 AR 模式参数的方法，我们令 $y_t,y_{t-1},\cdots,y_{t-p}$ 分别用 $y,y_1,y_2,\cdots,y_p$ 来代替。为此，可将(1)式改写如下：

$$y=\alpha_1 y_1+\alpha_2 y_2+\cdots+\alpha_p y_p+\varepsilon \quad (4)$$

式中下标 $1,2,\cdots,p$ 表示过去值。于是线性 AR 模式(4)的参数根据下列偏自相关剩余方差的关系

$$\left.\begin{aligned}\alpha_1^{(k)}&=\rho_{1y\cdot23\cdots k}\frac{\sigma_{y\cdot23\cdots k}}{\sigma_{1\cdot23\cdots k}}\\ \alpha_2^{(k)}&=\rho_{2y\cdot13\cdots k}\frac{\sigma_{y\cdot13\cdots k}}{\sigma_{2\cdot13\cdots k}}\\ &\cdots\cdots\\ \alpha_k^{(k)}&=\rho_{ky\cdot12\cdots k-1}\frac{\sigma_{y\cdot12\cdots k-1}}{\sigma_{k\cdot12\cdots k-1}}\end{aligned}\right\} \quad (5)$$

$$(k=1,2,\cdots,p)$$

来计算(Cramér,1946)。式中 $\rho_{1y\cdot23\cdots k}$ 称为 y 及 y_1 对 $y_2,\cdots,y_k$ 的偏自相关系数，$\sigma^2_{y\cdot23\cdots k}$ 称为 y 对 $y_2,\cdots,y_k$ 的剩余方差。其余那些 ρ 及 σ 的定义均相类似。

Draper 及 Smith(1967)在选择最佳多元回归模式时，曾用偏相关系数来描述前向挑选程序，然而在建立回归模式时却仍沿用 F 检验。作者(1977)提出一种藉助于将偏相关系数用于整个程序的 t 检验以建立多元回归模式，其结果与由逐步回归法得到的完全一样。相类似地，线性 AR 模式的参数可用(5)式予以估计，该模式的阶数则可由偏自相关系数的 t 检验

$$t=\frac{r_{ky\cdot12\cdots k-1}}{\sqrt{1-r^2_{ky\cdot12\cdots k-1}}}\nu^{1/2} \quad (6)$$

来选定。式中自由度 $\nu=n-1-k$，$r_{ky\cdot12\cdots k-1}$ 为偏自相关系数的估计值。(6)式与简单相关的 t 检验的形式相同，只不过这里其自由度为 $n-2-(k+1-2)=n-1-k$(von Mises, 1964, p. 608)。

在 AR 模式中，对于阶数越来越高的阶而言，(5)式中的逐步计算需要大量劳动。堪以告慰的是下列关系式

$$\left.\begin{aligned}&\sigma^2_{y\cdot12\cdots k-1}=\sigma^2_{1\cdot23\cdots k}=\sigma^2_{k\cdot12\cdots k-1}\\ \text{而}\quad&\\ &\sigma^2_{y\cdot23\cdots k}\neq\sigma^2_{y\cdot12\cdots k-1}=\sigma^2_{1\cdot23\cdots k}\\ &\sigma^2_{y\cdot13\cdots k}\neq\sigma^2_{y\cdot12\cdots k-1}=\sigma^2_{1\cdot23\cdots k}\\ &\cdots\cdots\cdots\cdots\cdots\cdots\end{aligned}\right\} \quad (7)$$

适用于线性 AR 模式。

所以，按(7)式可将(5)式中最后的方程表示为

$$\alpha_k^{(k)}=\rho_{ky\cdot12\cdots k-1},k=1,2,\cdots,p \quad (8)$$

此关系式指出如下事实：k 阶线性 AR 模式的最后参数 $\alpha_k^{(k)}$ 正好是后延 k 的偏相关系数(Box 及 Jenkins, 1970)。

例如，设 $k=1$，

$$\alpha_1^{(1)}=\rho_{1y}=\rho_1$$

设 $k=2$，

$$\alpha_2^{(2)}=\rho_{2y\cdot1}=\frac{\rho_{2y}-\rho_{12}\rho_{1y}}{[(1-\rho_{12}^2)(1-\rho_{1y}^2)]^{1/2}}=\frac{\rho_2-\alpha_1^{(1)}\rho_1}{1-\alpha_1^{(1)}\rho_1}$$

设 $k=3$，

$$\alpha_3^{(3)}=\rho_{3y\cdot12}=\frac{\rho_{3y\cdot1}-\rho_{23\cdot1}\rho_{2y\cdot1}}{[(1-\rho^2_{23\cdot1})(1-\rho^2_{2y\cdot1})]^{1/2}}$$

$$=\frac{(\rho_3-\rho_1\rho_2)-\rho_1(\rho_1\rho_3-\rho_2^2)+\rho_1(\rho_1^2-\rho_2)}{(1-\rho_1^2)-\rho_1(\rho_1-\rho_1\rho_2)+\rho_2(\rho_1^2-\rho_2)}$$

$$=\frac{\rho_3-[\alpha_1^{(2)}\rho_2+\alpha_2^{(2)}\rho_1]}{1-[\alpha_1^{(2)}\rho_1+\alpha_2^{(2)}\rho_2]}$$

上述结果正是得自(3)式的那些结果。

鉴于(8)式，为了选择线性 AR 模式的阶数，样本偏自相关系数可用(6)式予以检验。

将(8)式中的样本偏自相关系数代入(6)式可得

$$t=\frac{\hat{\alpha}_k^{(k)}}{[1-(\hat{\alpha}_k^{(k)})^2]^{1/2}}\nu^{1/2},k=1,2,\cdots,p \quad (9)$$

式中自由度 $\nu=n-1-k$，n 为用于计算后延 k 的偏自相关系数的观测记录的成对数目。所以，按预选的显著性水平利用 t 检验可以很容易选择线性 AR 模式的阶数，如果最后参数 $\alpha_k^{(k)}$ 在递推过程中的每一步都被估计出的话。为什么我们可用 t 检验来寻找具有内相关性的线性 AR 过程的阶数，这将在本文的最后予以详细证明。

已知 F 值为

$$F(1,n-k-1)=\frac{r^2_{y\cdot12\cdots k}-r^2_{y\cdot12\cdots k-1}}{1-r^2_{y\cdot12\cdots k}}(n-k-1) \quad (10)$$

式中 $r_{y\cdot12\cdots k}$ 为样本的多重相关系数，它在线性多元回归模式以及线性多元 AR 模式中选择预报因子程序内可作为一个检验统计量(见 Klein 等，1959；Buar, 1974, p. 385 关于线性多元模式的讨论)。如果在某种预选的显著性水平上，F 值超过了临界值 $F_{1-\alpha}$ 的话，那么新选的变量 y_k 可选为第 k 个预报因子，这样就建立了一个 k 阶的线性 AR 模式。

由下列关系式

$$r^2_{y\cdot12\cdots k}=1-(1-r^2_{1y})(1-r^2_{2y\cdot1})\times(1-r^2_{3y\cdot12})\cdots(1-r^2_{ky\cdot12\cdots k-1}) \quad (11)$$

(Kendall 及 Stuart, 1976, p. 355)，可以得到

$$F(1,n-k-1)=$$

$$=\frac{r^2_{y\cdot12\cdots k}-r^2_{y\cdot12\cdots k-1}}{1-r^2_{y\cdot12\cdots k}}(n-k-1)$$

$$=\frac{\begin{array}{l}(1-r^2_{y1})(1-r^2_{y2\cdot1})\cdots(1-r^2_{y(k-1)\cdot12\cdots k-2})-\\-(1-r^2_{y1})(1-r^2_{y2\cdot1})\cdots(1-r^2_{yk\cdot12\cdots k-1})\end{array}}{(1-r^2_{y1})(1-r^2_{y2\cdot1})\cdots(1-r^2_{yk\cdot12\cdots k-1})}\cdot(n-k-1)$$

$$=\frac{r^2_{yk\cdot12\cdots k-1}}{1-r^2_{yk\cdot12\cdots k-1}}(n-k-1)$$

所以，由(8)式

$$F(1,n-k-1)=\frac{[\hat{\alpha}_k^{(k)}]^2}{1-[\hat{\alpha}_k^{(k)}]^2}(n-k-1) \quad (12)$$

给出具有 1 及 $n-k-1$ 自由度的 F 分布。线性 AR 模式的阶数亦可用 F 检验来选定。所以，由(9)式及(12)式可得自由度 $\gamma_1=1$ 及 $\gamma_2=n-k-1$ 的如下表达式：

$$F(1,n-k-1)=t^2(n-k-1) \quad (13)$$

所以，如同(9)式及(12)式所指出的，将(3)式及(8)式结合起来以及或者用 t 检验或者用 F 检验，来选定 AR 模式的阶数是较为方便的，假设在递推程序(3)的每一步已经把最后参数都估计出来了的话。

3. 实际拟合

今以下个例阐述上述线性 AR 模式的拟合方法。表 1 给出上海四十年(1921—1960)六月份雨量记录的时间序列。此时间序列的前 30 个值用于拟合线性 AR 模式，后 10 个值则用于讨论不同阶数的模式的预报有效性。

为了表明上海这三十年(1921—1950)的时间序列是平稳的，不妨用下列方程

$$Y=\alpha+\beta t'+\varepsilon_{t'}$$

来拟合这些数据。在这里，α 和 β 为参数；$\varepsilon_{t'}$ 为随机变量；t' 为观测数，如果采用人为的时间原点，则 t' 为 $-5,-3,-1,1,3,5$ 等。我们发现

$$\left.\begin{array}{l}\hat{\alpha}=\overline{Y}=180.06,\\ \hat{\beta}=0.117,\ \mathrm{var}\,\hat{\beta}=0.939\\ \hat{\sigma}^2=8444,\ t=0.12\end{array}\right\}$$

可以这样说，这种趋势是很不显著的[$t\ll t_{0.95}(28)$]，由于此时间序列是平稳的，故可用以拟合线性 AR 模式。

利用上海三十年(1921—1950)六月份雨量记录的平稳时间序列，计算得到的自相关系数样本值如下：

$r_1=-0.339$　$r_2=-0.137$　$r_3=0.248$

$r_4=-0.0393$　$r_5=-0.116$　$r_6=0.0705$

表 1　上海六月份雨量 (mm)

年代	0	1	2	3	4	5	6	7	8	9
20		256.9	230.8	165.5	234.7	42.0	251.2	205.5	215.3	70.3
30	178.5	139.9	181.9	110.4	42.1	217.1	111.9	112.6	468.9	103.8
40	93.6	292.3	193.0	152.3	140.4	327.7	89.0	233.7	142.2	153.2
50	240.2	134.5	176.0	142.6	288.2	167.6	244.5	212.3	91.8	105.2
60	210.0									

作为第一步，由(3)、(9)及(12)式我们求得

$$\left.\begin{aligned}\hat{a}_1^{(1)}&=-0.34,\\ t&=-1.87, t<-t_{0.95}(27)\\ F&=3.51,\ F>F_{0.90}(1,27)\end{aligned}\right\}$$

此一阶线性 AR 模式的形式如下：

$$Y_t=241.11-0.34\,Y_{t-1} \qquad (I)$$

如果对于单尾检验采用 0.10 的显著性水平的话，则必须进行(3)式的计算。作为第二步，由(3)、(9)及(12)式我们求得

$$\left.\begin{aligned}\hat{a}_1^{(2)}&=-0.44, \hat{a}_2^{(2)}=-0.28\\ t&=-1.48,\ t<-t_{0.90}(25)\\ F&=2.20,\ \ F>F_{0.80}(1,25)\end{aligned}\right\}$$

于是，根据参数的显著性检验，二阶的线性 AR 模式可拟合为

$$Y_t=309.70-0.44\,Y_{t-1}-0.29\,Y_{t-2} \qquad (II)$$

这里，为了上海六月份雨量预报，Y_t 为预报量的现在的六月份雨量，Y_{t-1} 及 Y_{t-2} 为预报因子的过去两年的六月份雨量。

看一看三阶的线性 AR 模式是否可以拟合上海六月份雨量记录。为此，我们进行了如下计算

$$\left.\begin{aligned}\hat{a}_1^{(3)}&=-0.40, \hat{a}_2^{(3)}=-0.24, \hat{a}_3^{(3)}=0.11\\ t&=0.55,\ t>t_{0.70}(23)\\ F&=0.30, F<F_{0.50}(1,23)\end{aligned}\right\}$$

此三阶线性 AR 模式的形式如下：

$$Y_t=274.57-0.40\,Y_{t-1}-0.24\,Y_{t-2}+0.11\,Y_{t-3} \qquad (III)$$

根据单尾 t 检验的 0.10 显著性水平，上述三阶模式不能被选定。试再进一步，如果我们仅仅注意到拟合优度问题的话，则会设法对四阶模式，准确地说即

$$Y_t=251.64-0.41\,Y_{t-1}-0.22Y_{t-2}+0.15\,Y_{t-3}+0.08\,Y_{t-4} \qquad (IV)$$

作出鉴定。

我们都知道，倘若包括拟合资料和预报资料的整个序列是平稳的，并且在自相关中是遍历性的，那么，可预报性的提高取决于六月份雨量时间序列中的自相关的显著性。不妨这样推论，用考虑到拟合优度来鉴定而不用显著检验所鉴定的模式 IV，对预报效率要稍差一些。事实上，模式 IV 的 t 检验给出

$$t=0.38, t<t_{0.70}(21)$$

从而模式 IV 是不能采用的。

如果显著性水平 α 值小，那么放弃择一假设的概率往往是大的。所以当气象问题中并不着眼于检验的效率时，则往往将 α 的值选为 0.10，以期能构成一个卓有成效的检验结果（与 Thom 及 Thom 1972 年的结果相比较）。因而，如果根据 Thom 的工作我们采用 t 检验的 $\alpha=0.10$ 的显著性水平，或采用 F 检验的 $\alpha=0.20$ 的显著性水平，那么以长期预报为目的的模式 II 是颇为优越的。

为了选择线性 AR 模式的阶数，Akaike 计算了每一步最终预报误差的估计量

$$FPE_k=\hat{\sigma}_k^2[1+(k+1)/n], k=1,2,\cdots,p \qquad (14)$$

如果估计的线性 AR 模式是无偏的，则 FPE 就是独立样本平均均方误差的期望值。

剩余平方和的递推公式为

$$S_k=S_{k-1}[1-(\hat{a}_k^{(k)})^2] \qquad (15)$$

当拟合 k 阶线性 AR 模式时此均方差(Jones, 1975)为

$$\hat{\sigma}_k^2=S_k(n-k-1)^{-1} \qquad (16)$$

这就是说，$S_k(n-k-1)^{-1}$ 为 y 对 $y_1, y_2\cdots, y_k$ 的样本剩余方差。S_k 的初始值为 $S_0=(n-1)\hat{\sigma}^2$，而 $\hat{\sigma}^2$ 为 y_t 的无偏的样本方差。

Jones 的经验指出，拟合逐日资料的模式所有的阶数常常达到 25 或 50 这样的最大阶数。对于 FPE_k 为极小的阶数被选为最佳拟合。

将 Akaike 最终预报误差用于选择上海六月份雨量(表 1)的线性 AR 模式阶数时，我们根据 (15)、(16)及(14)式求得如下结果：

$$FPE_1=7975,\quad FPE_3=8279,$$
$$FPE_2=7838,\quad FPE_4=8801。$$

由于模式 II 的最终预报误差值为极小，故应该用它来表示上海六月份雨量的时间序列。根据 Akaike 方法得到的结论，与在这里用选择的参数一致。

Carr(1980)应用 Lorenz(1977)的研究结果，导出独立样本平均误差的两个新估计量，它们仅仅与下列样本统计量：

$$L_1=\frac{n-1}{(n-k-1)(n-k-2)}S_k \qquad (17)$$

$$L_2=\frac{n(n-1)}{(n-k-1)^2(n-k-1)}S_k \qquad (18)$$

$$(k=1,2,\cdots,p)$$

有关。在这里，L_1 用于有偏的剩余方差，而 L_2 则用于无偏的剩余方差。这两个估计量后来又用于 AR 模式(Carr, 1981)。

我们将 Carr 的判据用来确定上海六月份雨量(表 1)的线性 AR 模式的阶数。由(17)式及(18)式可得

$$L_1\begin{cases}L_{1,1}=8030\\L_{1,2}=7948\\L_{1,3}=8474\\L_{1,4}=9116\end{cases}$$

$$L_2\begin{cases}L_{2,1}=8297\\L_{2,2}=8504\\L_{2,3}=9041\\L_{2,4}=10501\end{cases}$$

在这里，L_1及 L_2的第二个下标表示模式的阶数。L_1的计算结果指出，必须将模式 II 当作最佳预报模式予以采用；得自于模式 II 的结果与由 FPE 或由 (9) 式给出的 t 检验所得结果相同。L_2 的计算结果表明，模式 I 可以是这样的最佳预报模式，它相应于 $\alpha=0.05$ 显著性水平的 t 检验所给出的模式。

为了全面检验拟合优度问题和预报有效性问题，我们用(15)式对上海三十年(1921—1950)六月份雨量记录分别计算了这四个模式的剩余平方和，并对另外十年 (1951—1960) 的记录用 (I)、(II)、(III)及 (IV) 分别计算其误差平方和。这些计算结果如表 2 所示。

表 2　剩余平方和、误差平方和及其它误差

		I	II	III	IV
剩余平方和 (1921—1950)		209343	192399	189925	188600
误差平方和 (1951—1960)		35749	38312	39734	41282
误差	极大	95.4	90.7	97.2	93.5
	极小	−104.8	−104.1	−110.2	−112.8
	较差	200.2	194.8	207.4	206.3

由表 2 可清楚看出：模式 IV 的剩余平方和的值最小；模式 I 的误差平方和的值最小，这一点与 Carr(1981)的结论相一致，即估计值 L_2在确定 AR 模式的阶数方面比 FPE 更好一些。然而，模式 II 给出的误差范围为最小，即对模式 II 而言，其个别误差的变化比模式 I 的为小。所以，在建立气候预报模式过程中，将 t 检验用于 $\alpha=0.10$ 显著性水平的情况能给出效率最好的检验的事实，将是值得特别注意的。

将 AR 模式拟合于平稳时间序列时，往往需要考虑这些剩余是否是随机的，以期证实该拟合模式能否提供一个资料适合的说明。在评论了用于分析 AR 过程剩余的各种统计工具，诸如 Durbin-Watson 检验、拟合不足检验 (Portmanteau lack of fit test)等以后，Chatfield(1980) 给出如下结论，即着重注意剩余相关 r_k 的前几个值，特别是对于后延 1，以便看出是否有一个比零大得多；假如只有一个 r_k 值恰好如此，尽管并没有足够的证据来摒弃此模式也就这样处理了。

对于模式 I，这些剩余的平均值为 −3.35，方差为 85.34^2；r_k 的前几个值为：

$$r_1=-0.1289,\quad r_2=-0.2133,\quad r_3=0.2324,$$
$$r_4=0.5069,\quad r_5=-0.1434。$$

对于模式 II，这些剩余的平均值为 −7.03，方差为 81.65^2；r_k 的前几个值为：

$$r_1=-0.0160,\quad r_2=0.0175,\quad r_3=0.153,$$
$$r_4=0.0422,\quad r_5=-0.0749。$$

在进行了 t 检验之后，我们发现模式 I 和模式 II 的剩余序列是彼此独立的变量(是具有 $r_k=0$ 的随机序列)。然而，必须指出如下事实：1)对于模式 II，此剩余序列接近于白噪音序列，其自相关系数 $\rho_k=0$；2)模式 II 在进行拟合方面有很大的改进；3) 对于模式 III 或模式 IV，它们的拟合改进得有限 (见表 2)；4) 文献中给出的有关独立样本均方误差的其它近似及估计量 (Carr, 1980)，都与 L_1 相似 (Carr, 1981)，而 L_1与这里所用的 t 检验或 F 检验程序却颇为吻合。所以，根据上述事实我们深信，为了说明这些数据，用 $\alpha=0.10$ 显著性水平的 t 检验或用 $\alpha=0.20$ 显著度水平的 F 检验所鉴定的模式 II 是更佳的。

诚然模式 II 在拟合上有很大改进，故据拟合 AR 模式中附加项的上述计算剩余平方和的方法，看来，由模式 I 计算的剩余序列中无疑至少存在一个系统性分量(Systematic component)。这是一个在 AR 模式中还存在滑动平均误差的问题(Kendall 及 Stuart, 1976, pp. 508—509; Katz 及 Skaggs, 1981)。这就等于拟合一个自相关滑动平均 (ARMA)模式到资料上，所以在该资料中仍有信息能被汲取。然而，ARMA 模式各参数的估计远远比 AR 模式的困难得多；为此，一旦较低阶的 AR 过程恰当地拟合时间序列，则无需乎考虑较为一般的 ARMA 过程了，正如由 Katz 及 Skaggs(1981)曾总结过的那样。其实，这正是作者宁愿采用模式 II 而不是模式 I (对模式 I 必须附加上滑动平均项)的另一个理由。

4. 自相关模式检验的假设中的自由度

人们曾指出：线性 AR 模式可用 t 检验或 F 检

验予以鉴定。但检验的假设只能严格用于独立正态随机变量，而气象上的时间序列却大都具有显著的短暂的内相关性。这种序列的相关性会减小有效自由度，结果使统计前提的检验低估了显著性。Nordφ (1966)曾对回归方程：

$$Y_{(t)}=\beta_0+\beta_1X_1(t)+\beta_2X_2(t)+\cdots+\beta_kX_k(t)+e(t) \tag{19}$$

详细地讨论过此问题。他证实所估计的剩余方差自由度的阶数近似等于

$$\nu=n-n^{-1}\sum_{j=0}^{k}\sum_{u,v=1}^{n}R_{u-v}\rho_{(j)u-v} \tag{20}$$

式中

$$R_{u-v}=\frac{E[e(u)e(v)]}{\sqrt{E[e^2(u)]E[e^2(v)]}} \tag{21}$$

是后延$(u-v)$时剩余的总体相关系数，且

$$\rho_{(j)u-v}=\frac{E[x_j(u)x_j(v)]}{\sqrt{E[x_j^2(u)]E[x_j^2(v)]}} \tag{22}$$

是由$(u-v)$的时间单位所分开的$x_j(t)$偏离值的总体相关系数；$x_0(t)$为常数。(20)式中的第二项反映了由内相关性引起的自由度部分。

如果时间序列用简单马尔可夫过程表示之，则(20)式为

$$\nu=n-\sum_{j=0}^{k}\frac{1+\rho_{(j)}R_1}{1-\rho_{(j)}R_1}+\frac{2}{n}\sum_{j=0}^{k}\rho_{(j)}R_1\frac{1-\rho_{(j)}^nR_1^n}{(1-\rho_{(j)}R_1)^2} \tag{23}$$

式中$\rho_{(j)}$为第一阶自相关系数，R_1为第一阶剩余自相关系数。

据 Mann-Weld 定理，我们现将 Nordφ 的上述结果用于线性 AR 模式。由于

$$\rho_{(0)}=1 \quad 且\ \rho_{(1)}=\rho_{(k)}=\rho_1$$

(23)式变为

$$\nu=n-\left[\frac{1+R_1}{1-R_1}+k\frac{1+\rho_1R_1}{1-\rho_1R_1}\right]+\frac{2}{n}\left[\frac{R_1(1-R_1^n)}{(1-R_1)^2}+k\rho_1R_1\frac{1-\rho_1^nR_1^n}{(1-\rho_1R_1)^2}\right] \tag{24}$$

$$(k=1,2,\cdots)$$

如果线性 AR 模式可用参数的 t 检验或 F 检验来鉴定的话，则(9)式中的自由度 ν 就应该以(24)式中的 ν 来代替。然而问题却在于怎样来估计剩余的总体自相关系数。k 值越大，则 R_1 的值越小。如果 k 值足够大，那么剩余将趋向于随机数目，以及 R_1 则趋近于零。

我们由(1)式可得

$$\begin{aligned}E[e^2(t)]=\sigma^2[&(\alpha_0^2+\alpha_1^2+\alpha_2^2+\cdots+\alpha_k^2)\\&+2\rho_1(-\alpha_0\alpha_1+\alpha_1\alpha_2+\alpha_2\alpha_3+\cdots+\alpha_{k-1}\alpha_k)\\&+2\rho_2(-\alpha_0\alpha_2+\alpha_1\alpha_3+\alpha_2\alpha_4+\cdots+\alpha_{k-2}\alpha_k)\\&+2\rho_3(-\alpha_0\alpha_3+\alpha_1\alpha_4+\alpha_2\alpha_5+\cdots+\alpha_{k-3}\alpha_k)\\&\cdots\cdots\\&+2\rho_k(-\alpha_0\alpha_k)]\end{aligned} \tag{25}$$

式中 $\alpha_0=1$；且

$$\begin{aligned}E[e(t)e(t-1)]=\sigma^2[&(\alpha_0^2+\alpha_1^2+\alpha_2^2+\cdots+\alpha_k^2)\rho_1\\&+(\alpha_1\alpha_2+\alpha_2\alpha_3+\cdots+\alpha_{k-1}\alpha_k)(\rho_0+\rho_2)\\&+\alpha_1\alpha_3(\rho_1+\rho_3)+\alpha_1\alpha_4(\rho_2+\rho_4)\\&+\cdots+\alpha_1\alpha_k(\rho_{k-2}+\rho_k)-\alpha_1(\rho_0+\rho_2)\\&-\alpha_2(\rho_1+\rho_3)-\cdots-\alpha_k(\rho_{k-1}+\rho_{k+1})]\end{aligned} \tag{26}$$

式中 $\rho_0=1$。

试设 $k=1$，

$$R_1=\frac{(1+\alpha_1^2)\rho_1-\alpha_1(1+\rho_2)}{(1+\alpha_1^2)-2\alpha_1\rho_1}$$

而当 $k=2$ 时，

$$R_1=[(1+\alpha_1^2+\alpha_2^2)\rho_1+\alpha_1\alpha_2(1+\rho_2)-\alpha_1(1+\rho_2)-\alpha_2(\rho_1+\rho_3)]/[1+\alpha_1^2+\alpha_2^2+2\rho_1(\alpha_1\alpha_2-\alpha_1)-2\alpha_2\rho_2] \tag{27}$$

对上海六月份雨量的线性 AR 模式 II，我们用 α 及 ρ 的样本值由 (27) 式及 (24) 式可求得下列结果

$$R_1=0.032 \qquad \nu=28-3.024$$

这些计算结果在很大程度上说明了这样的事实：由于模式 II 的 $\nu=28-3.024\approx25$，就具有给定的短暂的内相关性的月雨量的时间序列而言，t 检验或 F 检验可用来选择线性 AR 模式的适当阶数。自由度的这种值表明模式 II 很合适地建立在随机变量上面。

因为这种内相关性能有效地减小按(24)式给出的自由度，故由临界 t 值(得自表格形式的 t 分布中)所要求的自由度，可大于由(9)式求得的观测的 t 值。例如，假定 $r_1=0.3114$，以及 $t=1.7027$。但 95% 的临界 t 值对于 $\nu=27$，有 $t_{0.95}(27)=1.703$。由于 $t<t_{0.95}(27)$，在 5% 的显著性水平的情况下，我们应断定 $\rho_1=0$ 的零假定必然不能摒弃。因而，时间序列中有一种重要的自相关。可是假设在检验过程中独立随机变量是确实被满足的话，由于此自由度必须大于27，故观测的 t 值必须大于 1.7027。同时由表中找到的临界 t 值必然仍为 $t_{0.95}(27)=1.703$，因为 t 分布对于独立正态随机变量是能成立的。换句话说，时间序列中的这些短暂的内相关性将提高选择的阈

值；因而用减少自由度的 t 检验进行的前向选择，在判断显著性中是正确的，因为观测的 t 值仍然大于提高了的选择阈值。同时，在前向选择过程中，偏自相关系数由于时间间隔后延的增大而继续减小内相关性的影响。当某个不显著的参数最后在选择过程的某一步中被发现，则由 t 检验所要求的近似独立性的条件在确定线性 AR 模式的阶数时应被满足，有如以上计算结果表明的那样。

根据上述讨论，我们应该指出：在实际应用中，对于较大值的 $\hat{a}_k^{(k)}$ 进行 t 检验或 F 检验是不必要的，正如 Granger 及 Newbold 于 1977 年提出的逐步自回归那样（参见 Chatfield，1981），并且究竟我们能不能连续地计算（3）式这个问题是不能简单地确定的。所以 t 检验和 F 检验只是当 $\hat{a}_k^{(k)}$ 十分接近于零的时候才是必须的。这个程序与 Chatfield (1980，p.69)所完成的很相似。

赵颂华译自 Monthly Weather Review Vol. 111，No. 4，April 1983　宁宁校